Multi-omics Data Integration

Amrit Singh, PhD
Assistant Professor

June 7th, 2023 | 13:00-15:00 PST

TOG Intermediate Workshop
BCCHR Trainee Omics Group (TOG)

Comp Bio lab
code

THE UNIVERSITY OF BRITISH COLUMBIA
Department of Anesthesiology, Pharmacology & Therapeutics
Faculty of Medicine

Centre for Heart Lung Innovation
UBC and St. Paul's Hospital

CompBio Lab @ HLI
cbl-hli.med.ubc.ca/

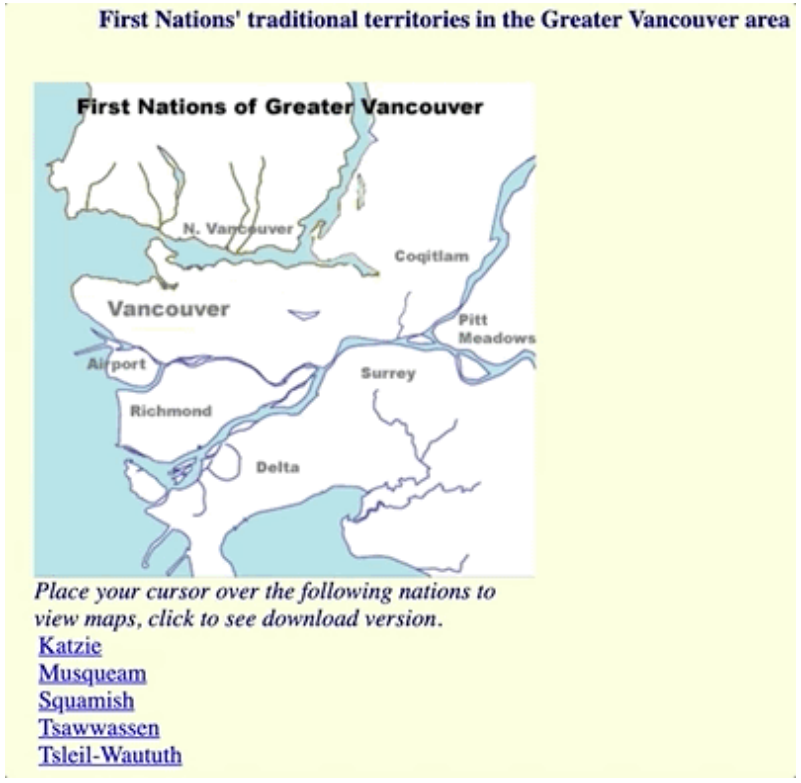
Land acknowledgement

I would like to acknowledge that I work on the traditional, ancestral, and unceded territory of the Coast Salish Peoples, including the territories of the xwməθkwəy̓əm (Musqueam), Sḵwx̱wú7mesh (Squamish), Stó:lō and Səl̓ílwətaʔ/Selilwitulh (Tsleil- Waututh) Nations.

Traditional: Traditionally used and/or occupied by Musqueam people

Ancestral: Recognizes land that is handed down from generation to generation

Unceded: Refers to land that was not turned over to the Crown (government) by a treaty or other agreement



What are your expectations from today's workshop?

Learning outcomes

By the end of this lecture you will be able to:

1. Describe what the *mixOmics* R-library can do.

2. Describe when to use which method and for what purpose (exploration, classification, integration).

3. Analyze data using mixOmics for various purposes (exploration, classification, integration)

High-dimensional data

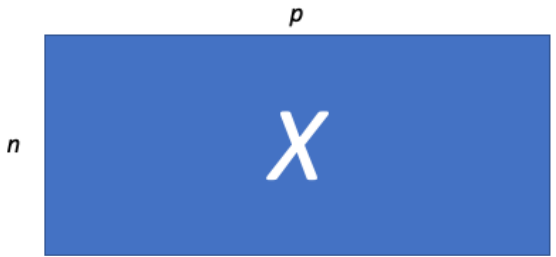
- n <<< p (number of observations is much smaller than the number variables)
- data is highly correlated

univariate

```
##        p_1
## 1  -0.6700
## 2  -0.0904
## 3   0.0200
## 4  -1.2000
## 5  -0.8140
## 6   0.8980
## 7  -0.1430
## 8   0.6100
## 9   1.2500
## 10  0.6200
```

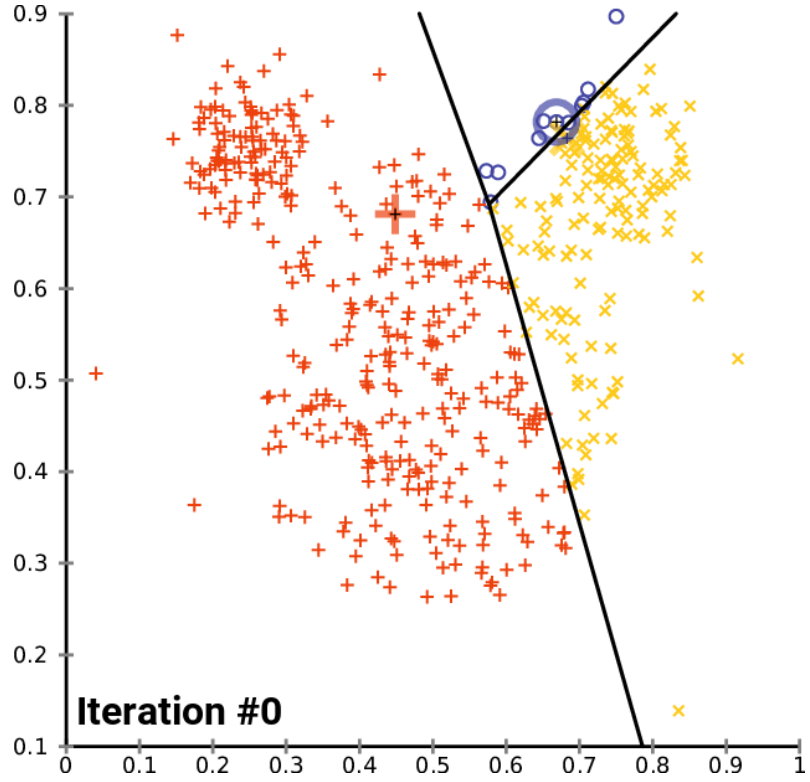
multivariate

```
##        p_1     p_2    p_3     p_4    p_5     p_6     p_7     p_8     p_9    p_10
## 1  -0.4990  1.6300  0.349 -0.8150 -1.090 -0.6810 -0.1820  0.6920  0.6080 -0.8450
## 2   0.9220  1.2200  1.140  0.0129 -0.568 -0.5780 -1.9100 -1.1700  0.4540 -2.2800
## 3   1.7100 -0.8570  0.662 -1.0500  0.819  2.4400  1.1600 -0.7990  0.8140 -1.3400
## 4  -1.4800 -1.5900  2.130 -0.1650  0.741  1.3800 -0.0738  0.1710 -0.2340  1.6800
## 5  -0.4970 -1.5300  0.939  0.0839 -0.238 -0.5450 -0.2970 -0.0242  0.0341 -0.3420
## 6   0.5310  0.0468  0.928 -0.0389 -0.913  0.0657  1.3000  0.0334  0.7830 -0.4860
## 7   1.1000  0.5110 -0.159  0.9280 -0.477 -2.8200 -1.5800 -0.0524  0.5650 -0.3160
## 8  -1.1000  0.4540  0.634 -0.6610  0.226  0.2930  1.2500 -1.5000 -0.0443  1.6000
## 9  -0.3690  2.3500 -1.740  0.0212  1.670 -0.3250 -0.7870  1.7900  0.5470  0.0477
## 10  0.0412 -1.7100  1.350  1.7700 -1.630  0.9670  0.0655 -1.1500  0.0854 -0.0550
```



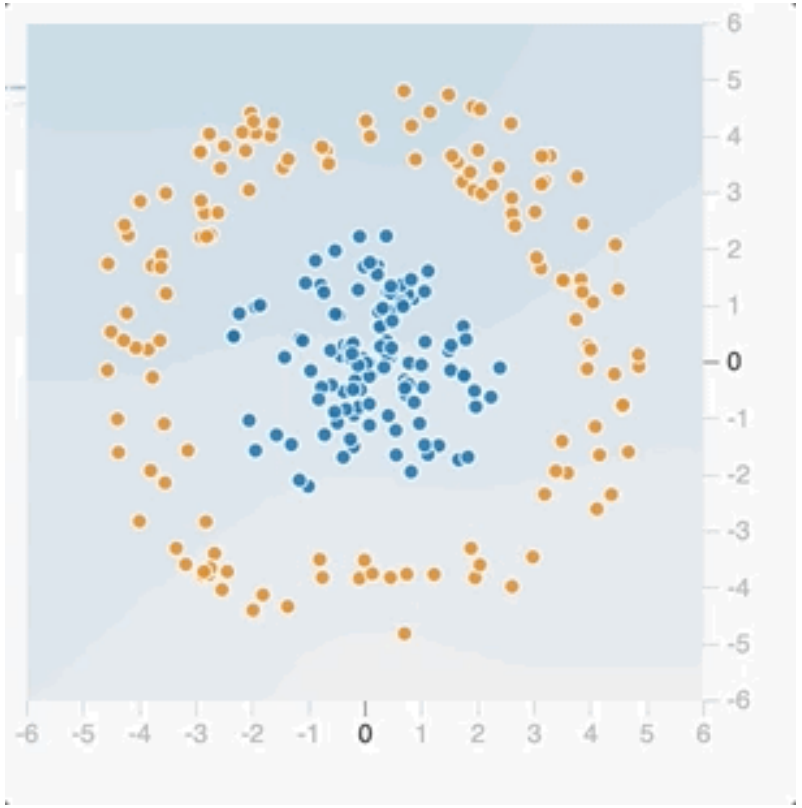
What can you do with high-dimensional data?

Unsupervised (clustering)



Chire 2017

Supervised (regression/classification)



Tensorflow playground

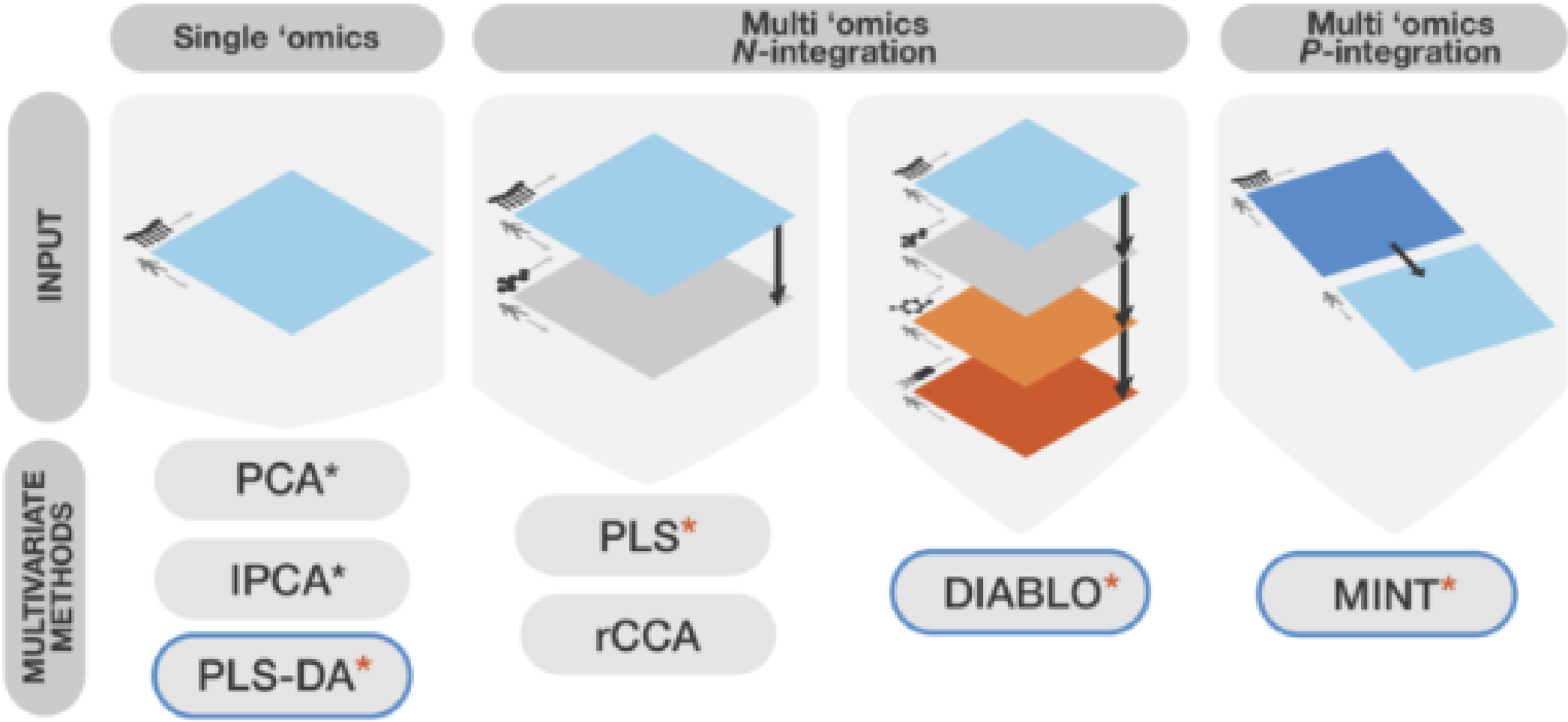
mixOmics

- initiative started and maintained by Prof Kim-Anh Lê Cao
- R-library with 19 methods for high-dimensional data (exploratory analyses, classification, regression, data integration, meta-analysis)

Lab head: A/Prof Kim-Anh Lê Cao
NHMRC Career Development Fellow
Melbourne Integrative Genomics (MIG) & School of Mathematics and Statistics
Building 184 ground floor | University of Melbourne | Parkville VIC 3010

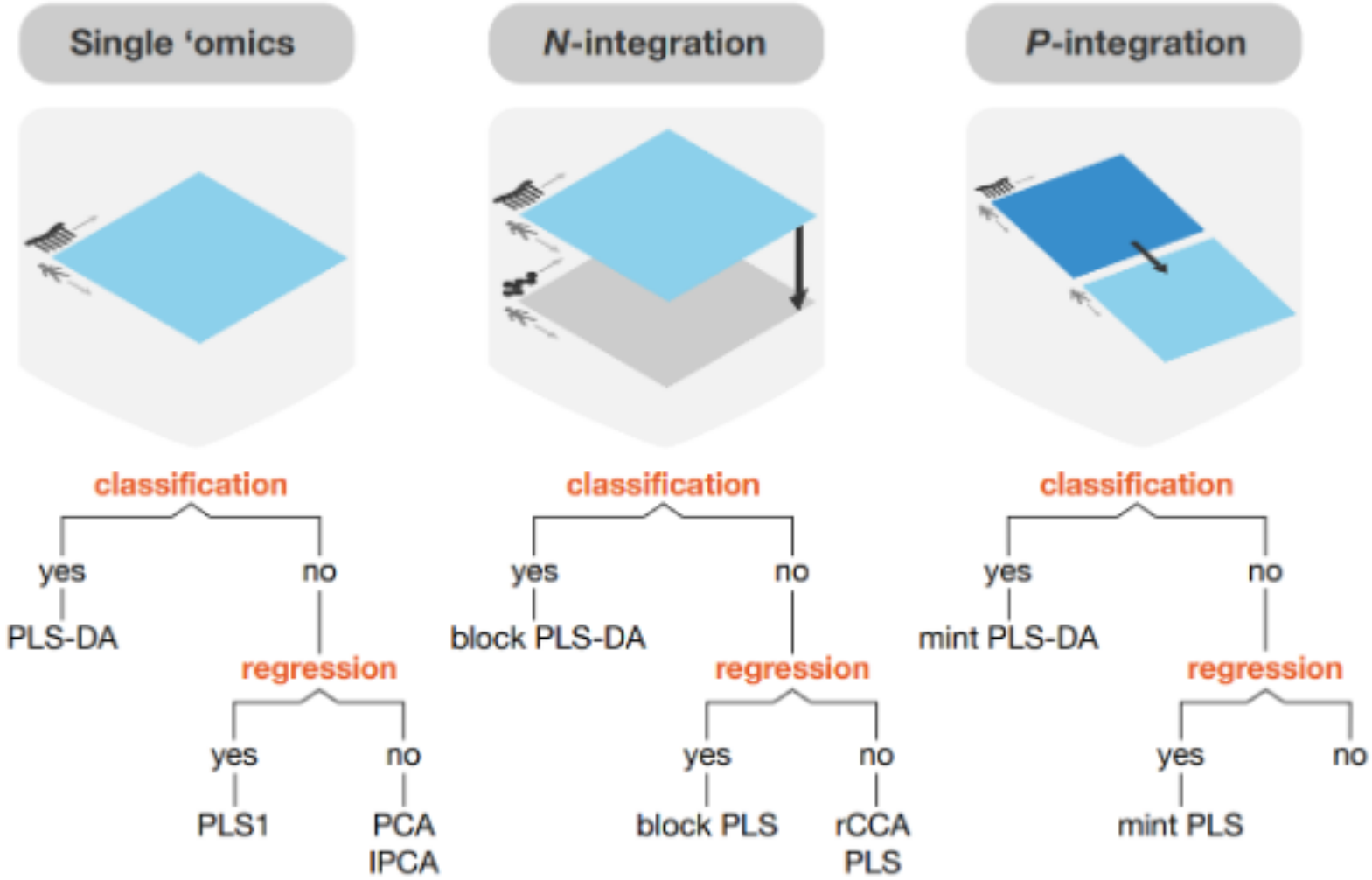
@: kimanh.lecao[at]unimelb.edu.au | twitter: @mixOmics_team | Ph: +61 3 8344 3971

What does mixOmics offer? methods...

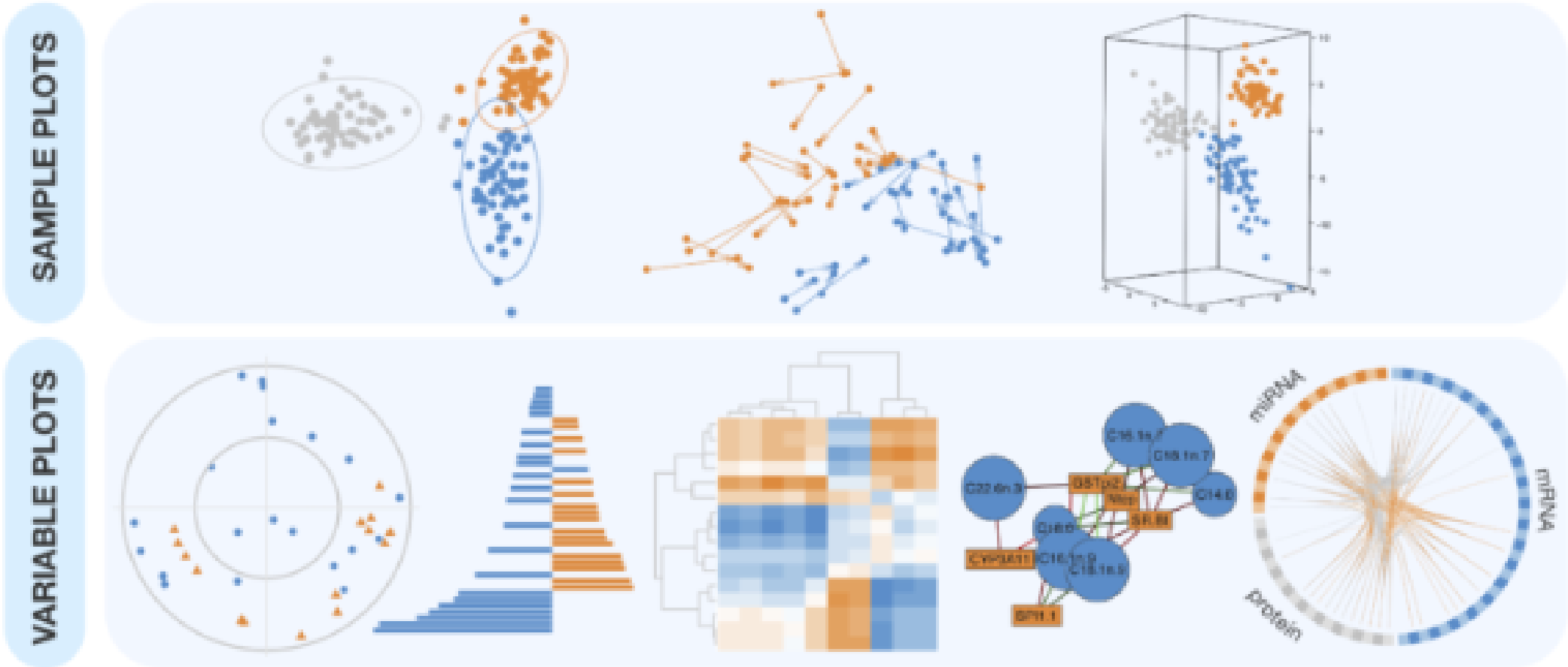


mixOmics.org | *variable selection

What does mixOmics offer? when to use these methods...



What does mixOmics offer? graphics...



mixOmics.org

Getting started with mixOmics

1. Download R
2. Download RStudio
3. install mixOmics

install mixOmics

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("mixOmics")
```

load vignette

```
openVignette("mixOmics")
```

Dataset used in this talk

Breast Cancer multi omics data from TCGA

This data set is a small subset of the full data set from The Cancer Genome Atlas that can be analysed with the DIABLO framework. It contains the expression or abundance of three matching omics data sets: mRNA, miRNA and proteomics for 150 breast cancer samples (Basal, Her2, Luminal A) in the training set, and 70 samples in the test set. The test set is missing the proteomics data set.

```
library(mixOmics)
data(breast.TCGA)
lapply(breast.TCGA$data.train, dim)
```

```
## $mirna
## [1] 150 184
##
## $mrna
## [1] 150 200
##
## $protein
## [1] 150 142
##
## $subtype
## NULL
```

breast cancer subtypes

```
addmargins(table(breast.TCGA$data.train$subtype))
```

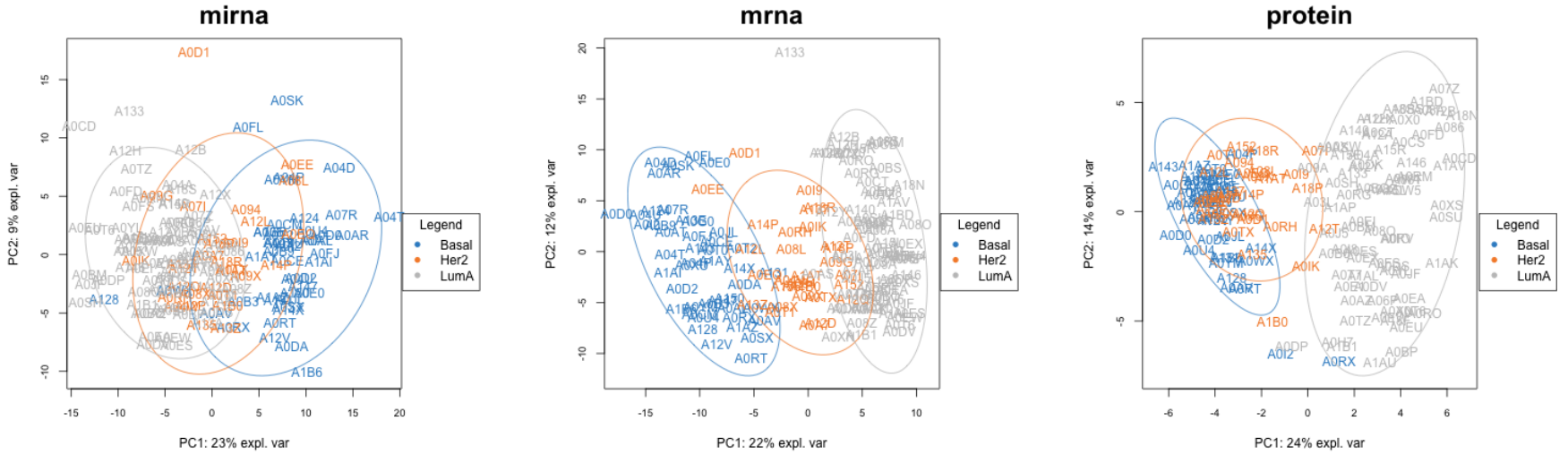
```
##
## Basal  Her2  LumA   Sum
##    45    30    75   150
```

Types of analyses covered:

Analysis	Methods	Functions	Input	Output
Exploratory data analysis	PCA	pca() plotIndiv()	X	
Discriminant analysis	sPLSDA	splsda() tune(), perf() plotIndiv(), plotVar()	X	Y
Data integration analysis	DIABLO	block.splsda() tune(), perf() plotDiablo(), circosPlot()	X_1, ..., X_J	Y

Exploratory data analysis using PCA

```
J <- length(breast.TCGA$data.train)-1
pcs <- lapply(breast.TCGA$data.train[1:J], pca)

mapply(function(pca, dataset){
  plotIndiv(pca,
            title=dataset,
            group=breast.TCGA$data.train$subtype,
            style="graphics",
            legend=TRUE,
            ellipse = TRUE,
            ellipse.level = 0.90)
}, pca=pcs, dataset=names(breast.TCGA$data.train)[1:J])
```



Discriminant analysis using sPLSDA

- based on the eda it seems **mrna** is better at separating classes than **mirna**, lets test this.
- this may or may not be true since we peeked at the data (need to test model with another dataset)

- mrna

```
mrna_model <- splsda(X = breast.TCGA$data.train$mrna,
                     Y = breast.TCGA$data.train$subty
                     keepX = c(5, 5),
                     ncomp = 2)
mrna_perf <- perf(mrna_model, validation = "Mfold", fo
mrna_perf$error.rate
```

```
## $overall
##         max.dist centroids.dist mahalanobis.dist
## comp1 0.22000000      0.1840000        0.1840000
## comp2 0.09733333      0.1026667        0.1413333
##
## $BER
##        max.dist centroids.dist mahalanobis.dist
## comp1 0.3496296      0.2035556        0.2035556
## comp2 0.1250370      0.1151111        0.1536296
```

- mirna

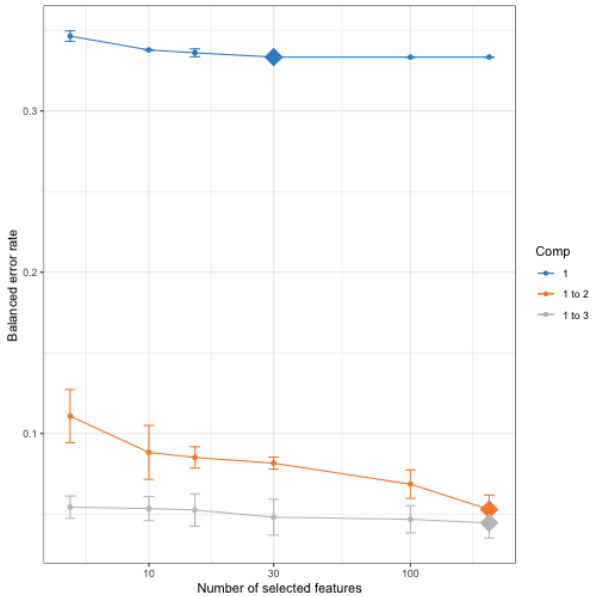
```
mirna_model <- splsda(X = breast.TCGA$data.train$mirna
                      Y = breast.TCGA$data.train$subty
                      keepX = c(5, 5),
                      ncomp = 2)
mirna_perf <- perf(mirna_model, validation = "Mfold",
mirna_perf$error.rate
```

```
## $overall
##        max.dist centroids.dist mahalanobis.dist
## comp1 0.2560000      0.3106667        0.3106667
## comp2 0.2133333      0.2266667        0.2120000
##
## $BER
##        max.dist centroids.dist mahalanobis.dist
## comp1 0.3819259      0.3400000         0.340000
## comp2 0.2930370      0.2451852         0.249037
```

How to select *ncomp* and *keepX*? use a grid of values

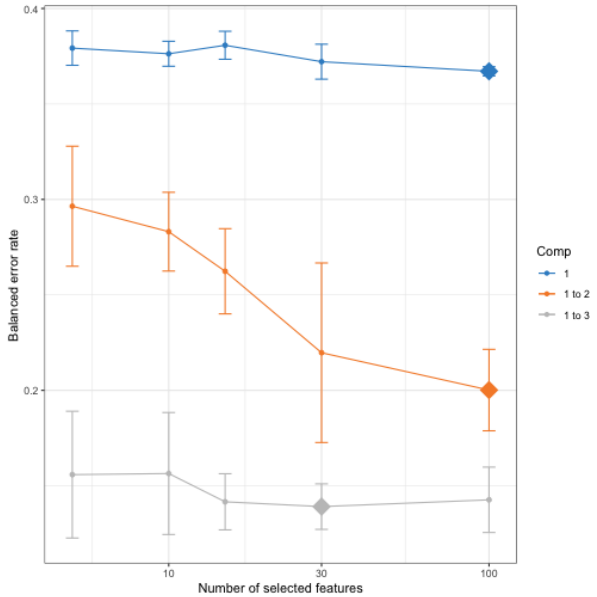
- mrna

```
tune_mrna = tune(method = "splsda", X = breast.TCGA$da
                 Y=breast.TCGA$data.train$subtype, nco
                 test.keepX = c(5, 10, 15, 30, 100, 20
                 progressBar = FALSE)
plot(tune_mrna)
```



- mirna

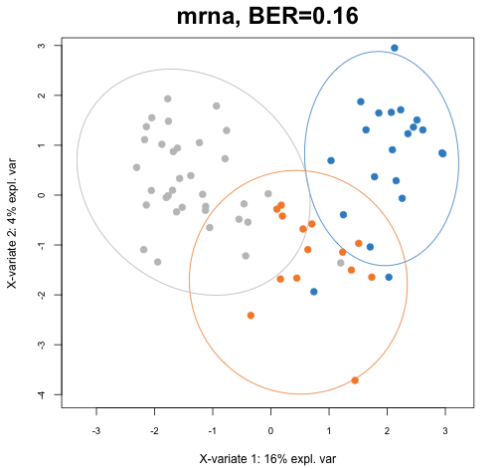
```
tune_mirna = tune(method = "splsda", X = breast.TCGA$d
                  Y=breast.TCGA$data.train$subtype, nco
                  test.keepX = c(5, 10, 15, 30, 100, 20
                  progressBar = FALSE)
plot(tune_mirna)
```



Test sPLSDA models using data from other observations (patients)

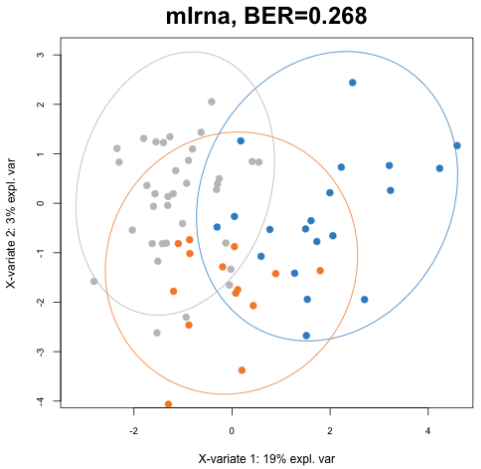
- mrna

```
mrna_model <- splsda(X = breast.TCGA$data.train$mrna,
                     Y = breast.TCGA$data.train$subty
                     keepX = rep(5, 3),
                     ncomp = 3)
mrna_pred <- predict(mrna_model, newdata = breast.TCGA

plotIndiv(mrna_model, comp = 1:2, rep.space = "X-varia
points(mrna_pred$variates[, 1], mrna_pred$variates[, 2
```



- mirna

```
mirna_model <- splsda(X = breast.TCGA$data.train$mirna
                      Y = breast.TCGA$data.train$subty
                      keepX = rep(5, 3),
                      ncomp = 3)
mirna_pred <- predict(mirna_model, newdata = breast.TC

plotIndiv(mirna_model, comp = 1:2, rep.space = "X-vari
points(mirna_pred$variates[, 1], mirna_pred$variates[,
```



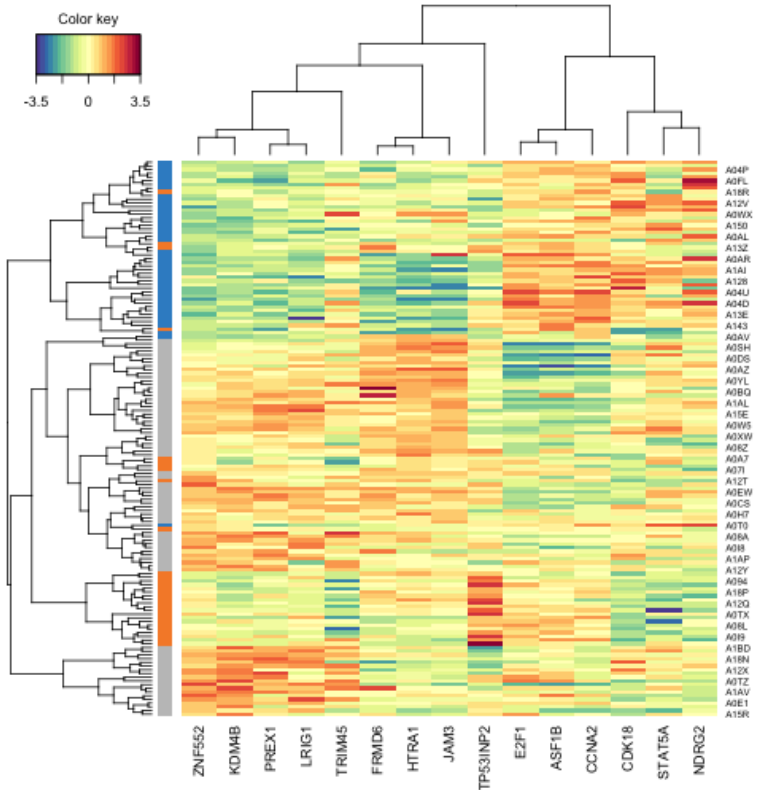
Model interpretation

- determine variables with most importance in mrna model

```
rbind(selectVar(mrna_model, comp=1)$value,
      selectVar(mrna_model, comp=2)$value,
      selectVar(mrna_model, comp=3)$value)
```

```
##             value.var
## ZNF552    -0.75801237
## KDM4B     -0.58296361
## PREX1     -0.20979766
## LRIG1     -0.17185790
## CCNA2      0.10963799
## CDK18      0.69045321
## TP53INP2  -0.68042479
## NDRG2      0.22678162
## STAT5A     0.07254351
## TRIM45     0.06003335
## JAM3       0.72727214
## E2F1      -0.53439117
## FRMD6      0.33920140
## ASF1B     -0.23142418
## HTRA1      0.12994836
```

```
cim(mrna_model,
    row.sideColors = mixOmics::color.mixo(as.numeric(b
```



DIABLO: an integrative classification method for multi-omics data

Design matters!

```
data = list(mrna = breast.TCGA$data.train$mrna, mirna
            protein = breast.TCGA$data.train$protein)
# set up a full design where every block is connected
# could also consider other weights, see our mixOmics
design = matrix(1, ncol = length(data), nrow = length(
                dimnames = list(names(data), names(dat
diag(design) =  0
design
```

```
##         mrna mirna protein
## mrna       0     1       1
## mirna      1     0       1
## protein    1     1       0
```

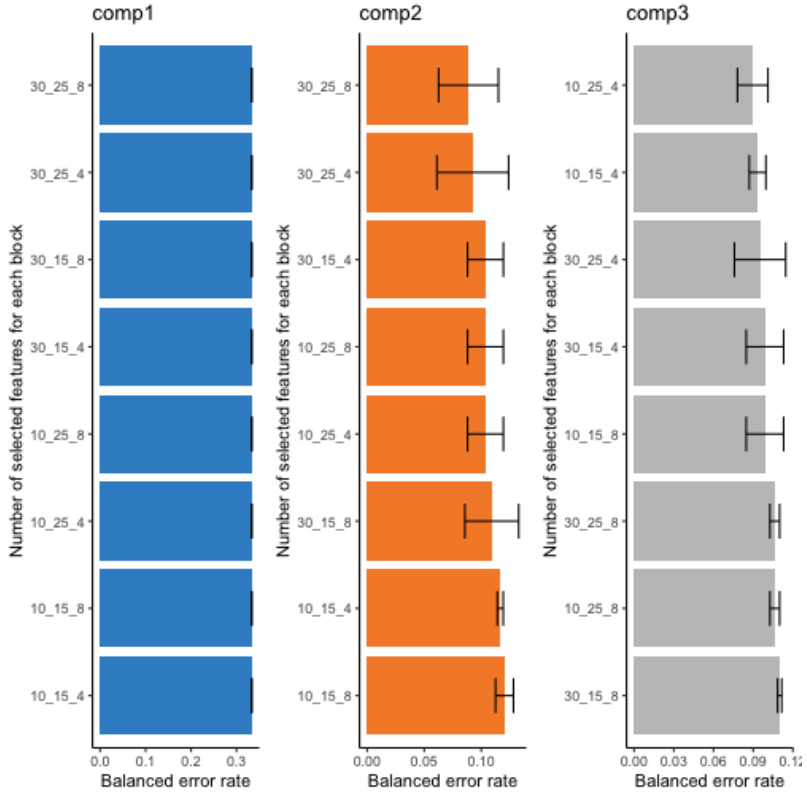
```
# set number of component per data set
ncomp = 3
test.keepX = list(mrna = c(10, 30), mirna = c(15, 25),

## setup cluster - use SnowParam() on Widnows
BPPARAM <- BiocParallel::MulticoreParam(workers = para
tune <- tune.block.splsda(
    X = data,
    Y = breast.TCGA$data.train$subtype,
    ncomp = ncomp,
    test.keepX = test.keepX,
    design = design,
    nrepeat = 2,
    BPPARAM = BPPARAM
)
```

```
## Design matrix has changed to include Y; each block will be
##              linked to Y.

##
## You have provided a sequence of keepX of length: 2 for block
## This results in 8 models being fitted for each component and
```

Bioinformatics. 2019 Sep 1;35(17):3055-3062.

Finding the optimal DIABLO model

```
plot(tune)
```

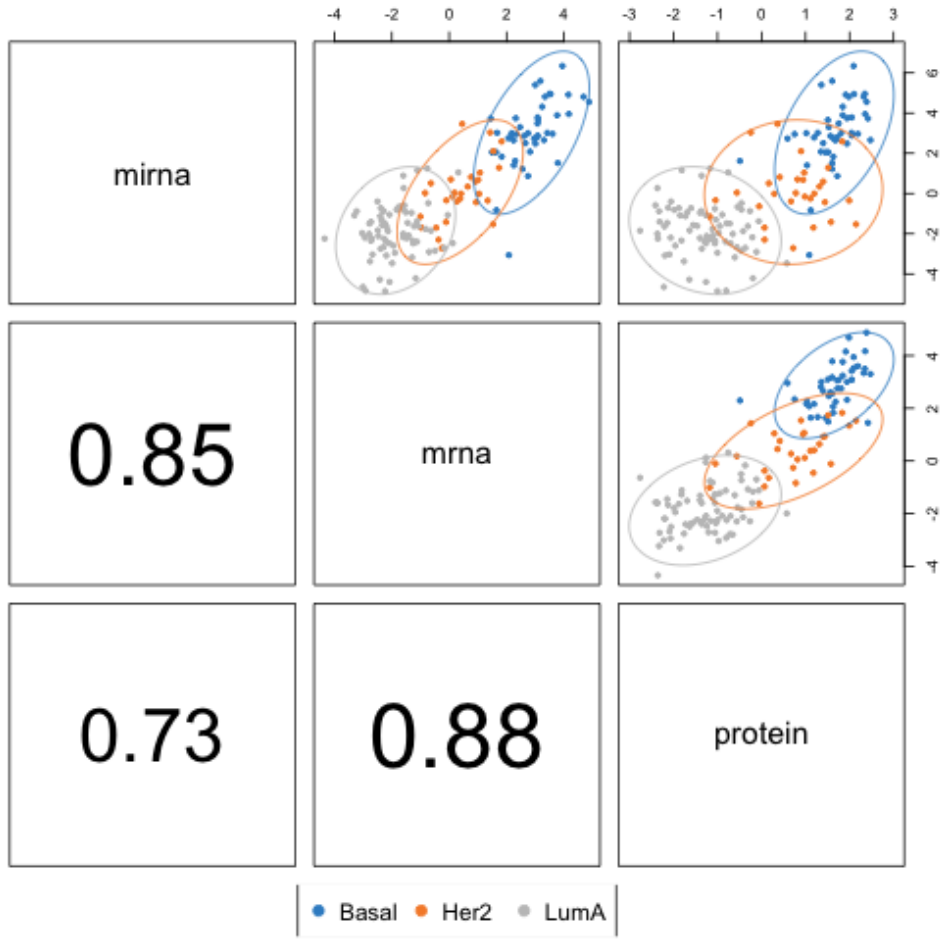



```
tune$choice.keepX
```

```
## $mrna
## [1] 10 30 10
##
## $mirna
## [1] 15 25 25
##
## $protein
## [1] 4 8 4
```

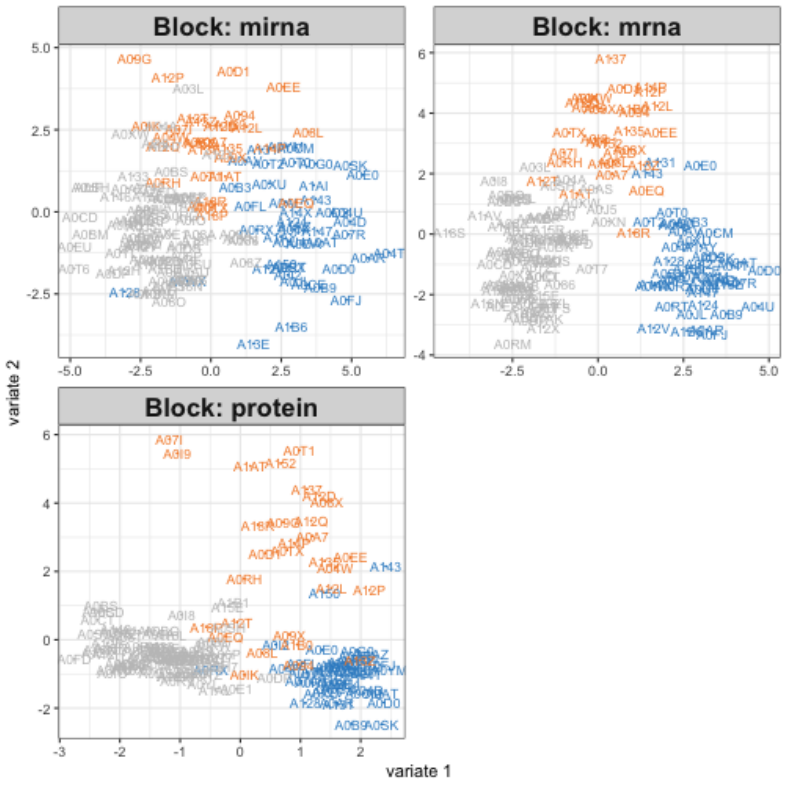
DIABLO model

```
ncomp = length(tune$choice.keepX$mrna)

diablo <- block.splsda(X = breast.TCGA$data.train[1:3]
                       Y = breast.TCGA$data.train$subt
                       keepX=tune$choice.keepX,
                       ncomp = ncomp)
```

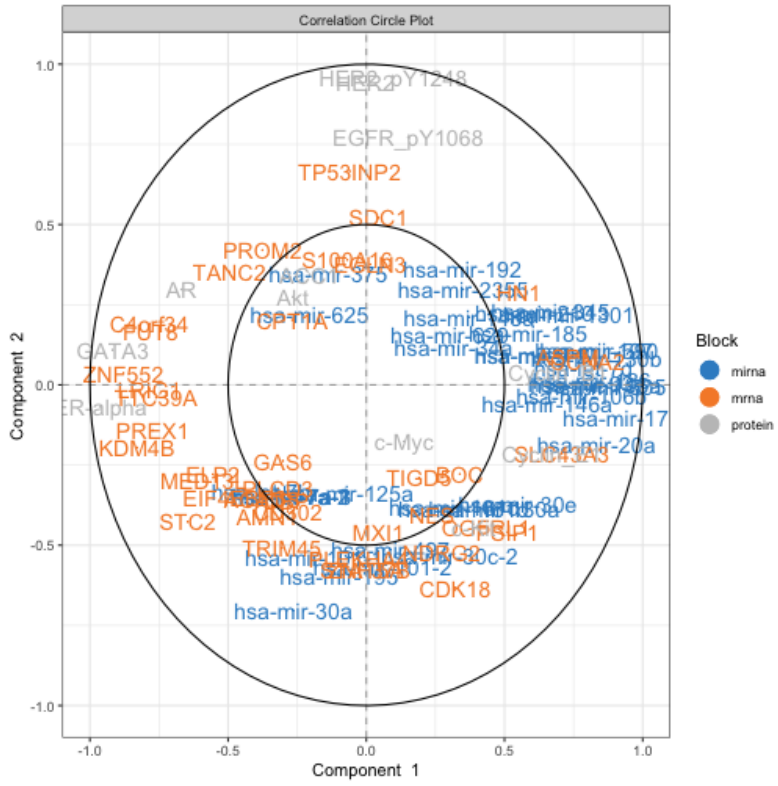


DIABLO: Sample and variable plots

```
plotIndiv(diablo)
```




```
plotVar(diablo, var.names = c(TRUE, TRUE, TRUE),
        legend=TRUE, pch=c(16,16,1))
```

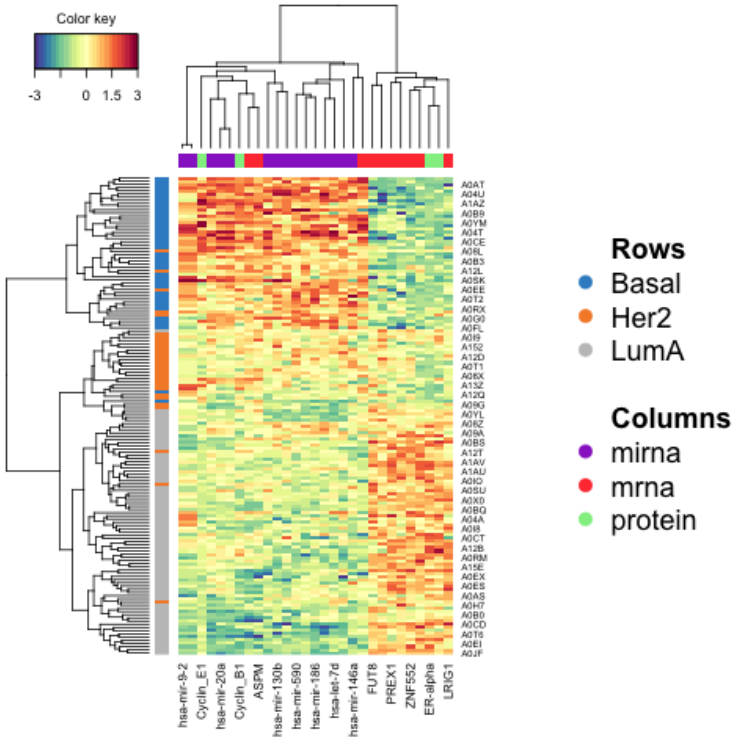


DIABLO

```
cimDiablo(diablo, color.blocks = c('darkorchid', 'brow
          comp = 1, margin=c(8,20), legend.position =
```

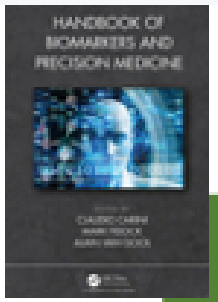
```
network(diablo, blocks = c(1,2,3),
        color.node = c('darkorchid', 'brown1', 'lightg
        cutoff = 0.75)
```

```
##
## trimming values to [-3, 3] range for cim visualisation. See 'trim' arg in ?cimDiablo
```



mikelove / awesome-multi-omics Public Watch 45 Fork 123 Star 502

seandavi / awesome-single-cell Public Watch 238 Fork 773 Star 2.2k



Unsupervised

Supervised

Data-driven

Knowledge-based

MCIA: omicade4
sMB-PLS[a]
SGCCA: RGCCA/mixOmics

Multiblock data analyses
Extensions of generalized CCA

SMSMA: msma
SGCCA+DA: mixOmics

SNF: SNFtool

Message-passing algorithms

PANDA: pandaR

Joint NMF[a]

Factorization methods

BCC: bayesCC

iBAG[b]

Bayesian Methods

Bayesian Networks[c]

Concatenation: caret
Ensemble: caretEnsemble

Classification and regression algorithms

SNMNMF[a]

Network-constrained

stSVM: netclass
GELnet: gelnet

Singh A et al., Handbook of Biomarkers and Precision Medicine CRC Press 2019:596

THE UNIVERSITY OF BRITISH COLUMBIA
Department of Anesthesiology, Pharmacology & Therapeutics
Faculty of Medicine

Centre for Heart Lung Innovation
UBC and St. Paul's Hospital

CompBio Lab @ HLI
cbl-hli.med.ubc.ca/

THANK YOU!

June 7th, 2023 | 13:00-15:00 PST

TOG Intermediate Workshop
BCCHR Trainee Omics Group (TOG)

Comp Bio lab
code